DOCUMENT RESUME

ED 266 633 FL 015 447

AUTHOR Klein-Braley, Christine
TITLE Reflections on the Content Validity of Holistic Tests.
PUB DATE 85
NOTE 11p.; In: Practice and Problems in Language Testing 8. Papers presented at the International Language Testing Symposium of the Interuniversitare Sprachtestgruppe (IUS) (8th, Tampere, Finland, November 17-18, 1984); see FL 015 442.
PUB TYPE Viewpoints (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS English (Second Language); Foreign Countries; *Holistic Approach; Language Proficiency; *Language Tests; Second Language Instruction; *Test Construction; Test Interpretation; *Test Validity
IDENTIFIERS *Content Validity

ABSTRACT

Traditional tests, sometimes called prescientific, and post-modern tests, such as cloze tests and dictations, take the same approach toward content validity: the examinee must show in an "integrated facile performance" that he or she can use the language fluently. Where traditional tests fail is in a third dimension, test method. These tests allow examinees to select their own language samples, with no guarantee that the samples are representative, as contrasted with post-modern tests that present a series of samples and require the examinee to demonstrate control of them. If it is found necessary to use traditional tests, some modifications should be made, including text sampling, stricter guidelines concerning the type of performance expected, and scoring procedures and guidelines for interpretation specified in advance. Because these modifications would require significant additional work, test users might consider using a post-modern test instead. A post-modern test might serve the teacher's purposes just as well as a traditional test since there is no difference in their approaches to content validity. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

BEST COPY AVAILABLE

REFLECTIONS ON THE CONTENT VALIDITY OF HOLISTIC TESTS

Christine Klein-Braley (University of Duisburg)

ED266633

Probably the test forms which cause the language tester the most headaches are those which need no justification whatsoever to the naive user: the essay, the translation, the précis, the written reproduction of an aurally presented text etc. In what Spolsky calls the traditional trend in language testing, such test procedures are the norm. Just as the budding pianist has to learn to read the music and practise scales in order to produce the final integrated performance of playing a sonata, so too such test forms offer the examinee the opportunity to demonstrate a synthesis of all that has been learned in a "facile integrated performance" (Carroll, 1961).

These traditional tests have high face validity. But I think that ultimately the validity concept underlying essays, translations etc. is that of content validity: I imagine that the majority of supporters of traditional tests would find it easy to agree with a statement like "holistic tests sample the learner's use of all elements of language simultaneously in an integrated fashion".

Content validity

The APA test standards state:

> Evidence of content validity is required when the test user wishes to estimate how an individual performs in the universe of situations the test is intended to represent. Content validity is most commonly evaluated for tests of skill or knowledge. (APA, 1974:26)

and continue:

> To demonstrate the content validity of a set of test scores, one must show that the behaviours demonstrated in testing constitute a representative sample of behaviors to be exhibited in a desired performance domain. Definitions of the performance domain, the user's objectives, and the method of sampling are crititical to claims of content validity. An investigations of content validity requires that the test developer or test user specify his objectives and carefully define the performance domain in light of those objectives. The definition should ordinarily specify the results of learning rather than the processes by which learning is either acquired or demonstrated. It should be sufficiently detailed and organized to show the degree to which component tasks make up the total domain. (APA, 1974:28)

FL 015 447

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

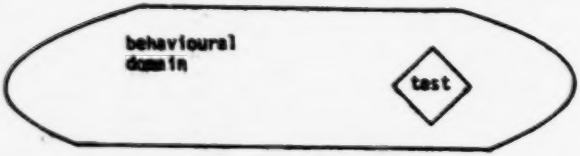
J. Tramola

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Graphically, we can represent this type of validity by Figure 1. The test is a representative sample of the behavioural domain, and therefore conclusions can be drawn from test performance to real-life performance.

Figure 1

Concept of content validity



The critical terms in the APA definition are:

- definitions of the performance domain
- the user's objectives
- the method of sampling.

Here I should like to discuss the implications of modern approaches towards content validity for traditional language tests.

Approaches to content validity

It is reasonable easy to ensure content validity when the domain to be sampled is finite and when all possible test items in the domain can be listed, even if only theoretically. When I was teaching infants our teaching objective in number skills at the end of the first year was the "number bonds to 20". This meant that a child had to be able to add, subtract, multiply and divide any two numbers to obtain a third, but that no term in the operations was to be bigger than 20. You can see that it is very easy to determine whether any given item belongs to the domain or not, and also that it is possible to list every item which belongs to the domain. However this would be rather a long list, so it is easier to write a rule which will generate suitable items. This might run:

$$a \begin{matrix} + \\ \times \end{matrix} b = c \text{ where } a, b, c \leq 20$$

The test would then consist of a random sample of such items, where the terms a & b are generated randomly. From test performance we could draw generalisations about whether a child had mastered the number bonds to 20 or not.

However there is an additional problem. The items have to be made

BEST COPY AVAILABLE

into test items, i.e. one of the terms a, b, or c has to be removed so that the child can demonstrate his or her mastery of the operations by inserting the correct solution. My experience of teaching infants tells me, that in fact the items are easier if term c is omitted

i.e. $4 \times 2 = ?$

than if a or b is omitted:

i.e. $? \times 2 = 8$
$4 \times ? = 8$

It is easy to see why this is the case. In the last two items the child has to work out that he or she must perform not a multiplication operation as the arithmetic operator implies, but a division if the correct solution is to be found. So omission of a or b demands a higher level of understanding of the operations concerned than omission of a. In other words, it is not only the representative sampling of the domain which is important, but also the types of item involved. This means that test method too is an integral part of the total test procedure.

The simplest technique for ensuring content validity is:

1. to define the domain either by listing the items or by devising rules which generate only items which belong to the domain;
2. to take a random sample from the domain;
3. to select a suitable item form for the testing.

Obviously, as in all tests, the items should be objective and equally obviously more than one item should be used in order to minimise measuring error. The final test is valid as a result of the construction procedures: content validity is built in right from the start.

In most cases the problem is not as easy to solve as in our example. For instance the domain can consist of an infinite number of items, or it can be rather difficult to determine where the domain boundaries are since it, itself, is infinite. For this reason it has become usual nowadays to construct tests intended to have content validity by means of what is generally called a Tyler matrix.

The Tyler matrix

The Tyler matrix is a two dimensional matrix with behavioural classes on one axis and content classes on the other:

BEST COPY AVAILABLE

For our example, the Tyler matrix would look like this:

Figure 2

Tyler matrix

CONTENT

Behaviour	Solve a op b = ?	Solve ? op b = c Solve a op ? = c
+		
-		
x		
÷		

The behavioural classes are usually hierarchically ordered, so that satisfactory performance in a higher-order category automatically includes the lower-order category.

Content validity and psychometric-structuralist tests

If we now turn our attention to one of the language testing classics, Harris 1969, we find that language testing too has long been using such matrices:

Figure 3

Language test construction matrix

	Language Skills			
Components	Listening	Speaking	Reading	Writing
Phonology/ Orthography				
Structure				
Vocabulary				
Rate & gen. fluency				

(Harris, 1969 : 1oo)

You see, therefore, that psychometric-structuralist language tests too are based on the concept of content validity.

BEST COPY AVAILABLE

If we take the Harris matrix as a basis for test construction , it is easy to see that in fact it is a supermatrix. Each of its cells, or each row can be enlarged into a separate matrix. For instance, we might choose to select the row STRUCTURE, and to subdivide this into a further set of content classes:

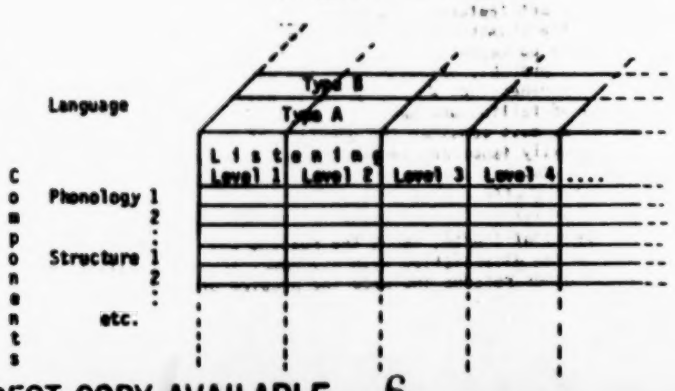
STRUCTURE

Determiners
Prepositions
Tenses
Active / Passive
.
.
.

or to subdivide LISTENING into categories which begin with the relatively easy listening task of a text, presented straight, read with "teaching pronounciation" so that every word can be clearly heard, with no hesitations, back tracks or "erms", and no background noise and which end with the most difficult task: listening to a genuine, unscripted conversation between native speakers, with all the "erms", interruptions, structural changes in mid-sentence and background noise that is typical for real-life communication. These categories or levels of difficulty are clearly hierarchical in a way that the original Harris categories are not.

The content classes in this case could either be those of the original Harris matrix, maybe broken down into subcategories, or we could choose to use a sampling procedure for all the various types of text content which are possible: a lecture on engineering, a scripted discussion on women's lib., an unscripted family discussion on where to spend the holidays this year. These must be regarded only as samples, as examples of possible subject matter, because one could just easily choose a lecture on extra-sensory perception, a scripted discussion on whether the 16-year-old daughter should be allowed to go to a pyjama party or not. Or we could try to include both Harris's components and the various topic areas in the sampling scheme. This would give us a three-dimensional matrix:

Figure 4



BEST COPY AVAILABLE

This too, is a familiar matrix to language testers: it is the one proposed by Cooper in 1968.

We still have the question of test method, i.e. item construction. We can test listening comprehension by using individual open-ended or multiple-choice items, we can ask the examinee to make deductions from what he or she has heard, to give us a brief summary. This is rather difficult to represent graphically, and so we use the mapping sentence (cf. Klauer, 1984). For our listening comprehension matrix the sentence would state something like:

The examinee demonstrates listening comprehension ... by correctly solving	ITEMS Type 1 . . . n	..which test..	COMPONENTS Phonology 1 . n ... 1 . n	of..	
....	DISCOURSE Type 1 2 . . n	at ...	LISTENING Level 1 2 . . n		

If we want to make deductions about the entire matrix we can either sample each cell once, which would give us

$$n_{(items)} \times n_{(components)} \times n_{(discourse)} \times n_{(listening\ levels)}$$

or we can take a random sample of the cells. Of course, if we want more information about any specific cell, we have to take a number of items from that cell. As you can see, test length is rapidly becoming infinite at this point.

There is another problem too. So far I have assumed that it is fairly easy to specify the substructures of the various dimensions, but, as we all know, this has not yet proved to be possible. Both the structuralists and the Chomskyans made brave attempts, but both ignored important features of the language: the structuralists its creativity, the Chomskyans its functionality. So the components are going to cause us headaches. So far as discourse topics are concerned, the number is probably infinite, and the same goes for levels of listening comprehension. We are going to have to sample here, but we have no way of telling whether our samples are representative or not. In addition we must consider the question of weighting. Is each of the cells equally important in the total activity of listening comprehension, or are some more important than others, and if so, which?

These were, of course, among the reasons why in the early seventies linguists became dissatisfied with psychometric-structuralist testing procedures. Other reasons included the problem of relating test beha-

BEST COPY AVAILABLE

viour to real-life language performance, i.e. of predicting the ability to function linguistically under real-time constraints from tests whose artificiality was only too apparent. Yet another reason was provided by the statistics emerging from psychometric-structuralist tests: the subtests developed for such allegedly distinct skills as listening and reading comprehension, grammar and vocabulary, all correlated so highly with each other that there seemed good reason to assume that, in fact, language was a unidimensional entity, and that language proficiency had to be viewed as holistic.

As a result of the apparent impossibility of ensuring content validity in psychometric-structuralist tests new test forms were proposed, among others dictations, cloze tests and later C-Tests. What about the content validity of these tests?

Content validity and post-modern tests

Post-modern tests such as cloze tests, dictations and C-Tests are all based on the theory that language is a holistic entity, and that the same type of language processing occurs whether the activity is "active" or "passive". For this reason the behavioural dimension of the matrix can be collapsed into one single column, headed "all". The former content specification also collapses into one dimension, because now the behaviour involves the use of all elements specified here as part of the language processing activity. This means that the content dimension can now be specified as "discourse topic: any", because post-modern tests make the assumption of text equivalence.

Figure 5

Content	Behaviour
	Language processing (all elements)
discourse topic: any	POST-MODERN TESTS

Obviously the same arguments can be advanced for all the traditional tests: in the essay we also have the assumption that all aspects of language processing are involved, and that the discourse topic is irrelevant. The same applies to such tasks as translations, summaries, interviews, and so on. Thus, if cloze tests, C-Tests, and dictations are content valid according to this approach, so too are the traditional test forms.

However, we also have the third dimension of the matrix, namely test method, and here there are major differences.

Text selection

So far as test construction is concerned, virtually all methods begin with text selection: one text. The text is viewed as a random sample of the universe of texts. However, as Munnally points out (1967), every test item has a component of specifity: characteristics which are unique to any particular item. Item specifity is undesirable,

BEST COPY AVAILABLE

because in any test it is not performance on the individual item which interests us; the item is seen as one of many possible items, i.e. again it is a sample. The more item specifity an item has, the less suitable it is as a test item. In holistic tests the use of only one text means that item specifity can play a much larger role than is desirable. Item specifity is at its most problematical in the essay, of course, because the instruction "Write an essay of 300 words on the following topic..." is a direct invitation to the examinee to produce his or her own text. The 5 or 6 texts of the C-Test are a much better approximation to a sampling process.

Test production

This stage is only relevant for post-modern tests. While the traditional tests simply take the selected text and ask the examinee to perform according to the instructions given, post-modern tests use the concept of random sampling again to delete segments of the text for examinee processing. In cloze tests this is the "nth word" deletion principle, in C-Tests the "rule of 2". In both cases the systematic deletions are intended as an approximation of random sampling. As I have shown, "nth word" deletion is not entirely satisfactory (Klein-Braley 1981), while the "rule of 2" does delete representative elements of the text. But the intention of sampling is present in both cases.

In the dictation random sampling is not necessary because here the examinee is confronted with the entire text and is asked to convert the aural input into a graphic representation.

The importance of this step lies in its effect on the scoring procedures. In all the post-modern tests there is a correct solution to the test item involved. Admittedly one can choose whether to count spelling errors or not, but in dictations it is usual to deduct one point for every wrong word, in cloze and C-Tests to add one point for every correct restoration. In cloze tests we have the added problem of whether to use exact scoring or acceptable scoring. Exact scoring is entirely objective; acceptable scoring can be made objective if the examiners are prepared to spend sufficient time over it.

Traditional tests, on the other hand, are not objective. In translations we have the notorious problem of deciding on which type of fidelity we are aiming for. Often this is phrased as "as free as necessary but as accurate as possible. So it is quite conceivable that two scorers would disagree entirely on the evaluation of an examinee's rendering of the text, even though it was linguistically correct.

Once again the essay is the most problematical test form, since not only do we have the problem of scoring examinee production, we even have to devise some technique for bringing entirely different texts (those produced by the individual examinees) into the same scale of measurement. Consider the two extremes: one examinee writes very short, very simple sentences, all of which are linguistically correct. A second examinee is more ambitious, writes more complex sentences, but misses the target some of the time. Who gets the higher mark?

BEST COPY AVAILABLE

The important of the test production techniques in post-modern tests is that they ensure objectivity at the same time as performing sampling.

Text (test) processing

In working through the text the examinee presents the examiner with a sample of his language. In post-modern tests this sample is determined by the text. Avoidance behaviour is not possible; the examinee must cope with what is set in front of him. This means that we are more likely to get a representative sample of the examinee's performance than in the traditional test forms where within broad limits the examinee decides what he or she will offer to the examiner. This is most obvious in the essay: both what the examinee says and the way in which he or she says it are a matter of free choice, and it would be a particularly stupid examinee who deliberately chose to use structures known to cause problems or to venture into content areas where the vocabulary is not in the repertoire. This also applies to the rather more structured tasks such as translation and précis where the source text controls the content but the linguistic formulations are left to the examinee. In fact, a large proportion of my teaching time is devoted to showing examination candidates how to "get round" problem areas in their examinations.

Test interpretation

Because they are objective, post-modern tests provide us with a numerical score at the end of the test procedure. This performance score is considered to be a numerical representation of the examinee's competence. This assumption is legitimate, because we have used sampling processes to produce the test. Naturally we need to take factors such as test reliability and the error of measurement into consideration, but in general we can assume that the examinee which the higher score is actually "better" at the language, i.e. his or her competence is more highly developed.

It is very difficult to make such assumptions on the basis of traditional tests, since the process of converting the scorer's evaluations of examinee performance into a numerical scale is a very tricky one for obvious reasons. This means that test interpretation is very difficult.

Conclusions

Traditional tests (or what Spolsky formerly called prescientific language tests) and post-modern tests both take the same approach towards content validity. The examinee is supposed to show in an "integrated facile performance" that he can use the language fluently.

Where traditional tests fail is in the third dimension of the matrix: test method. The post-modern tests present the examinees with a sample or samples of the language and expect them to demonstrate their control of the samples. By doing this they have a reasonable chance of obtaining a representative sample of examinee performance. Traditional tests, however, allow examinees to select their own samples, and so there is no guarantee that the test scripts are representative.

BEST COPY AVAILABLE

If it is considered necessary to use traditional tests, then there are a number of recommendations which can be made. They all involve more work, but this is necessary in order to tighten up the test process:

1. More than one text should be used (text sampling)
2. Examinees should be provided with much stricter guidelines about the type of performance expected (test construction and text processing)
3. Scoring procedures should be specified in advance and techniques for converting assessment into numbers should be developed, together with guidelines for interpretation (test interpretation).

In view of the additional effort these suggestions would involve, test users should consider whether a post-modern test would not serve their purposes just as well as a traditional test since so far as their approaches to content validity are concerned there are no differences.

Note

I have been greatly helped in the writing of this paper by my discussions with Ulrich Raatz. Any errors are, however, all my own work.

Bibliography

APA: American Psychological Association, 1974. Standards for educational and psychological tests. Washington: APA.

Carroll, J.B. 1961. "Fundamental considerations in testing for English language proficiency of foreign students." In Testing the English proficiency of foreign students. Washington: Center for Applied Linguistics. 30-40.

Cooper, R.L. 1968. "An elaborated language testing model." In Brown, H.D. (ed) 1968. Language Learning Special Issue No. 2: 57-72.

Harris, D.P. 1969. Testing English as a second language. New York: McGraw-Hill.

Klauer, K.-J. 1984. "Kontentvalidität." Diagnostica 30, 1: 1-23.

Nunnally, J.C. 1967. Psychometric theory. New York: McGraw-Hill.

BEST COPY AVAILABLE